A Modern Machine Translation Parable: the Linguistically Savvy Tortoise and the Hare Who Only Knew How To Count

(The Wascally Wabbit *Always* Wins)

Kishore Papineni

IBM T.J. Watson Research Center

Yorktown Heights, NY

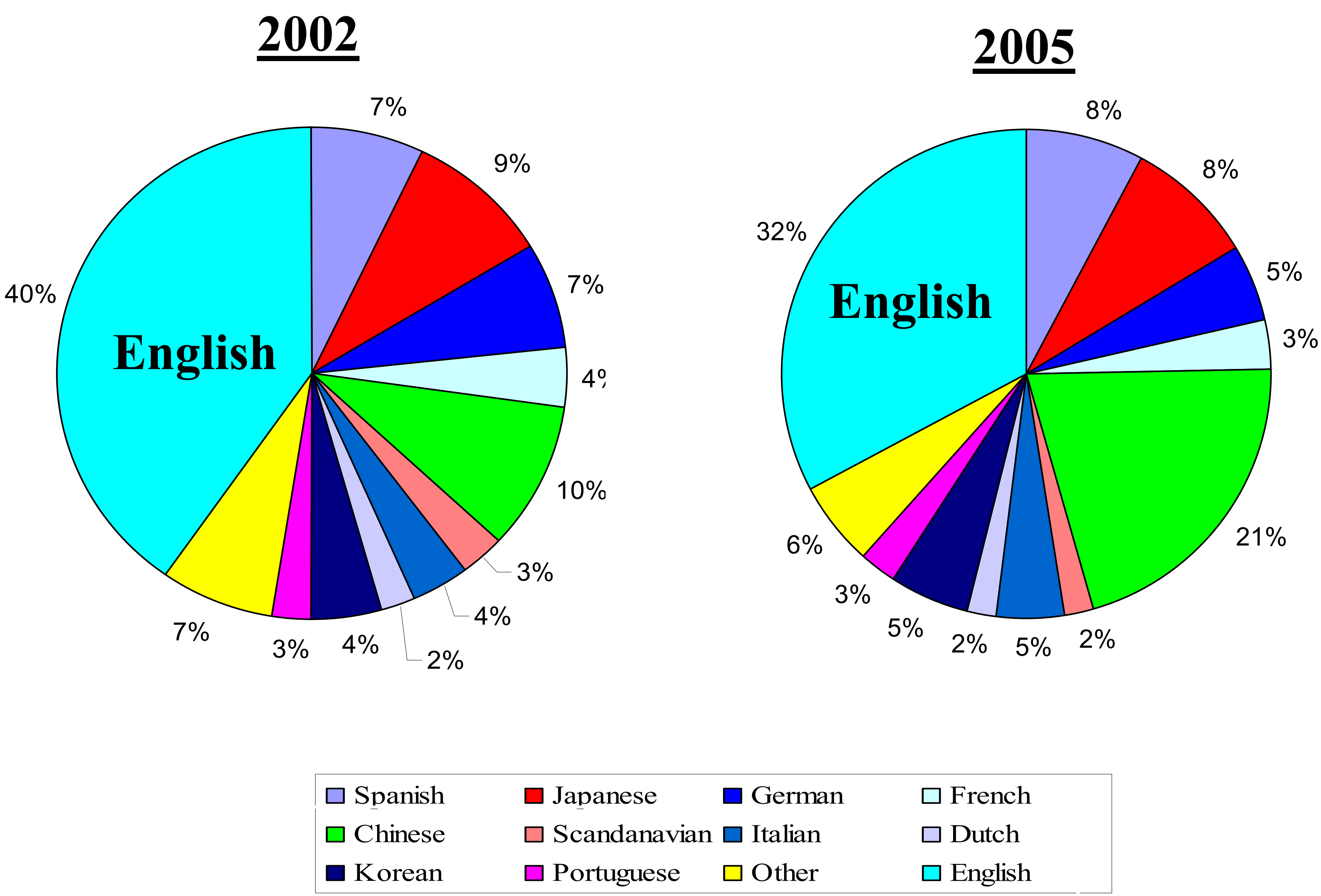
What is Machine Translation?

Automatically translate source language text into target language.

Why Machine Translation?

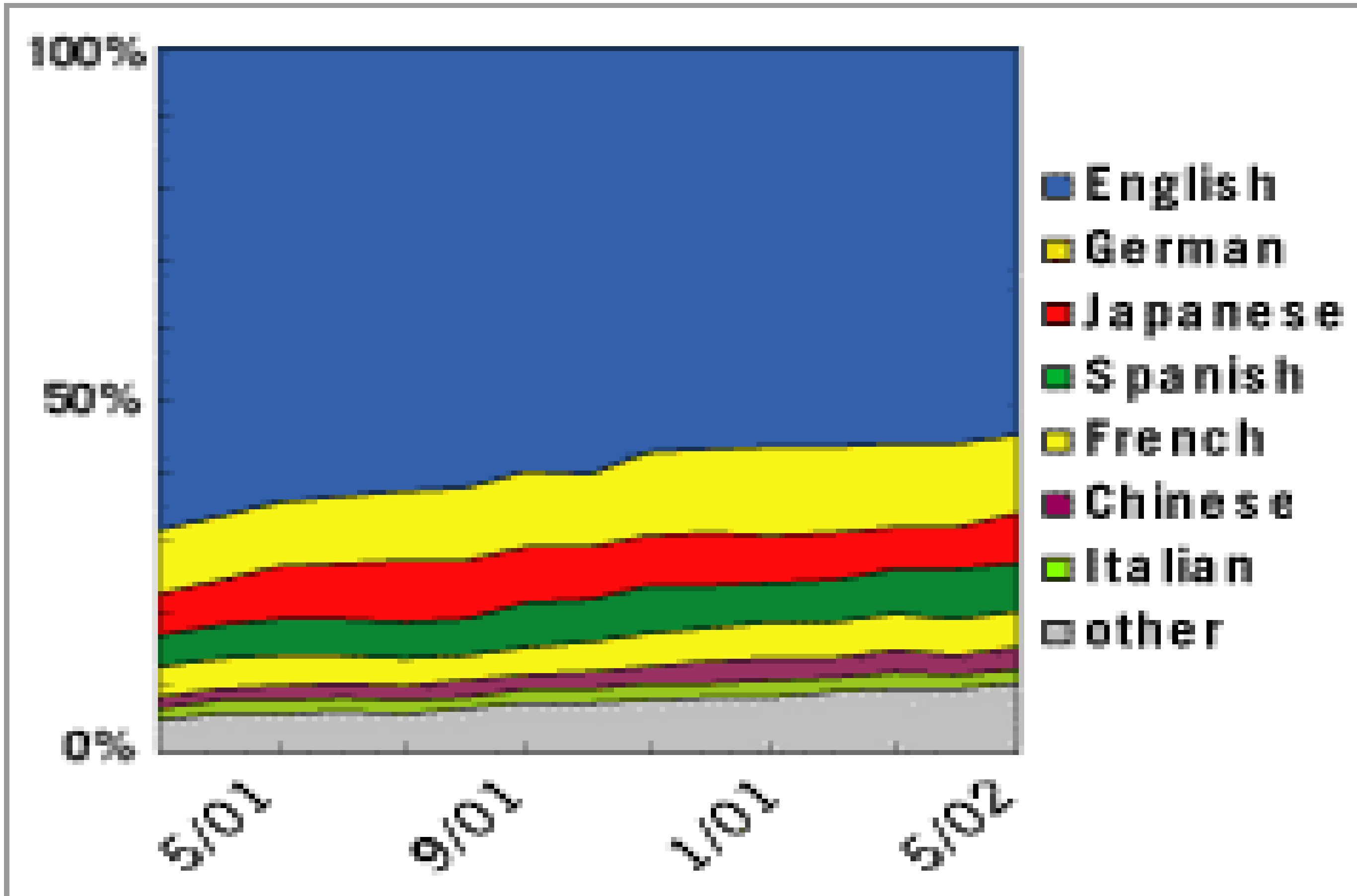
This is an exciting new era for MT!

Global Internet User Population



Languages Used to Access Google

English down to 55% from 70% in a year

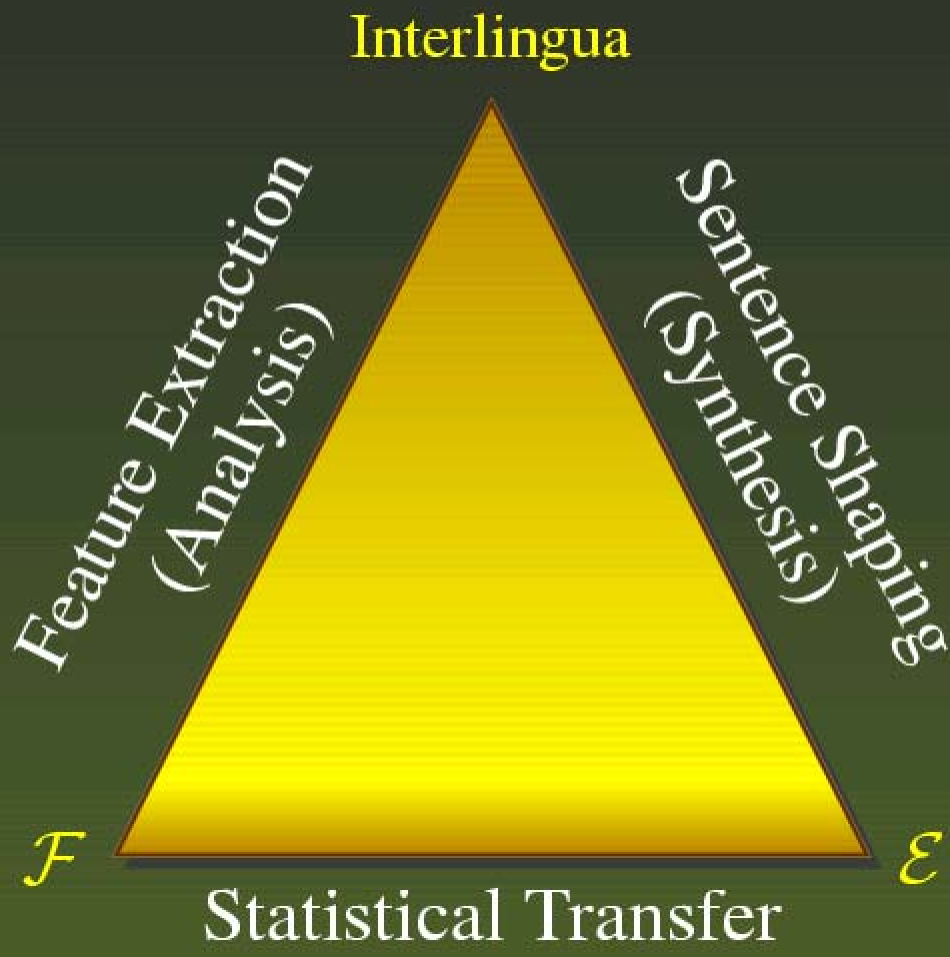


Cross-lingual Information Access

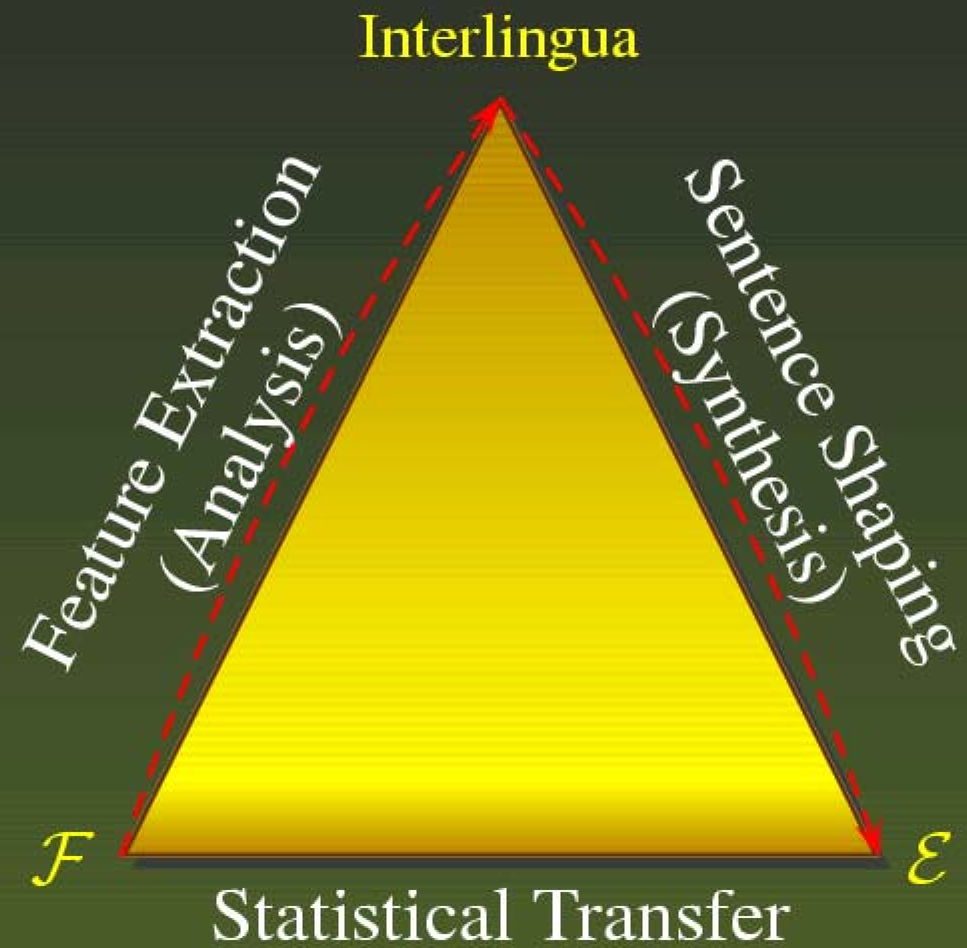
- Creates big demand for high-quality MT
- Is fueled in turn by high-quality MT

Worldwide resurgence of MT research

The MT Triangle

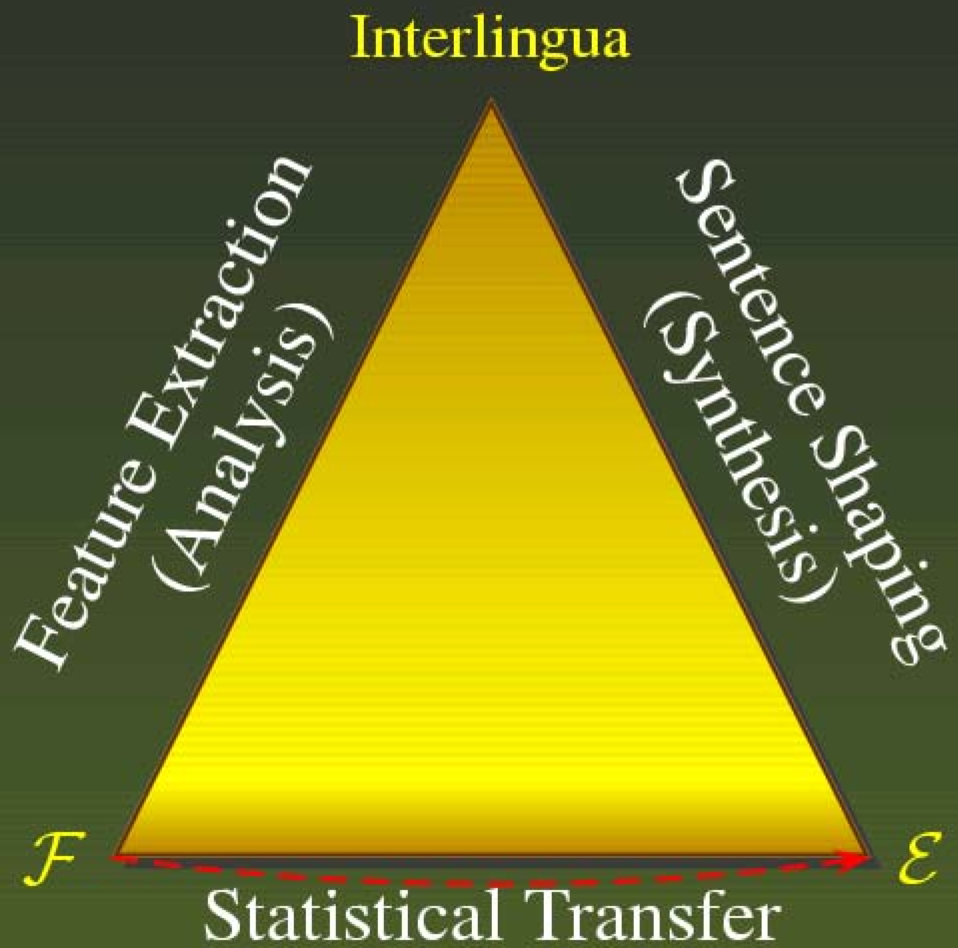


The MT Triangle



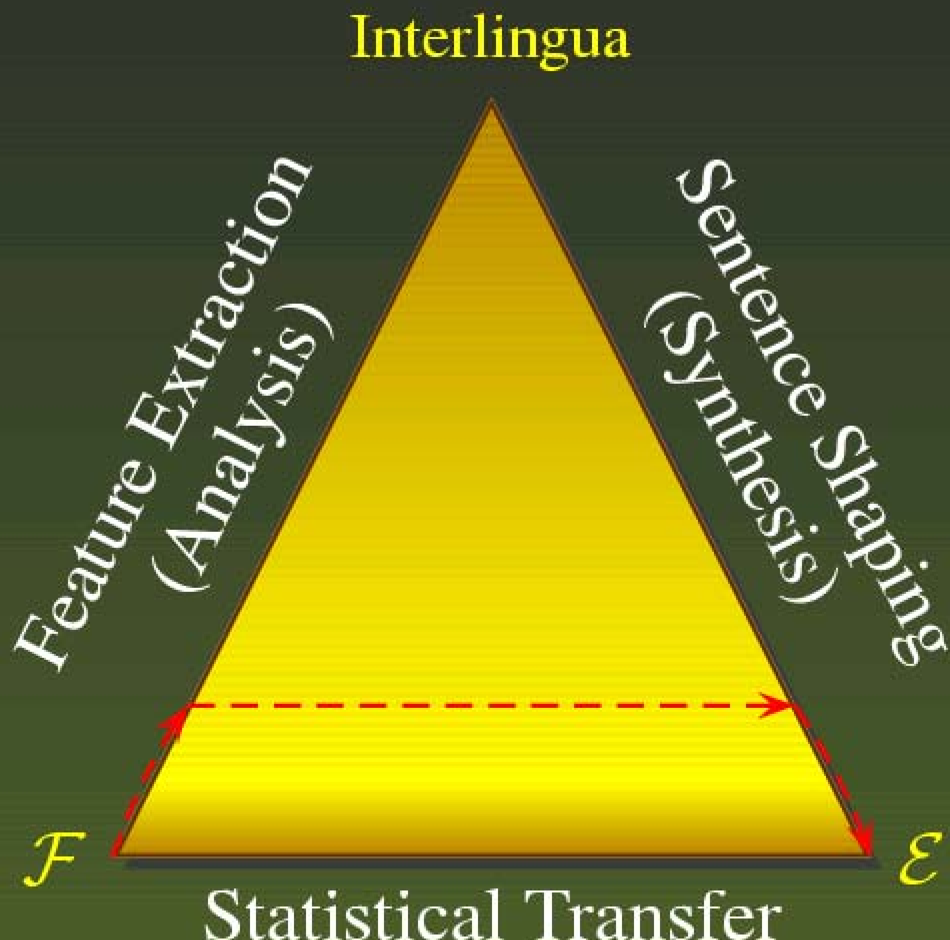
Linguistic high road all the way to interlingua and back

The MT Triangle



Pure statistical transfer

The MT Triangle



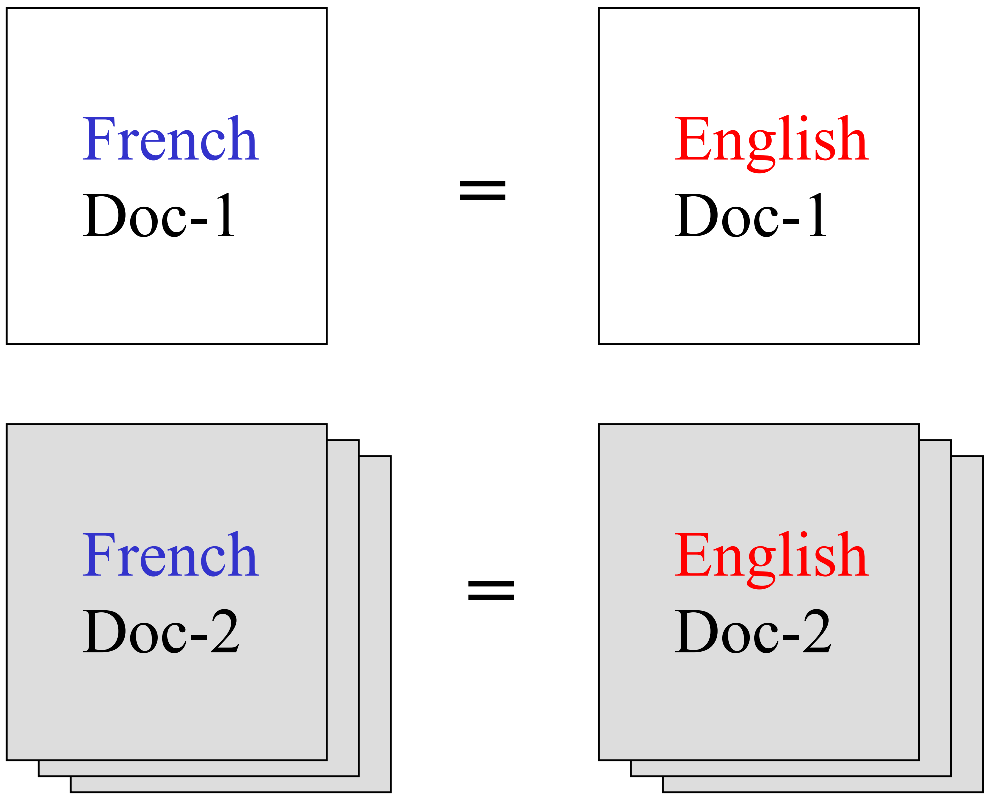
Practical engineering combination

Theory-based or Corpus-based?

Rule-based MT: write rules for translation based on grammar, manual dictionaries, ..

Statistical MT: automatically learn to translate from a parallel corpus of human translations

Parallel Corpus of Human Translations



Parallel Corpus - refined

E: The House met at 2 p.m.

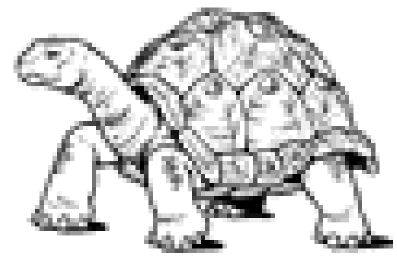
F: La séance est ouverte à 2 heures.

E: Mr. Speaker, I rise on a question of privilege affecting the rights and prerogatives of parliamentary committees and one which reflects on the word of two ministers of the Crown.

F: Monsieur l'Orateur, je soulève la question de privilège à propos des droits et des prérogatives des comités parlementaires et pour mettre en doute les propos de deux ministres de la Couronne.

Rule-based MT

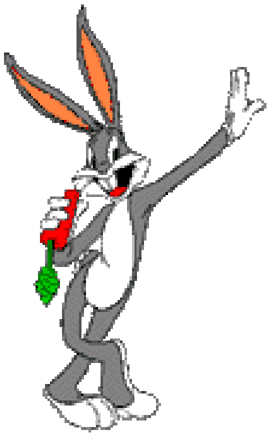
- Requires human expertise
- Expensive
- Slow: takes years to develop



- Not proven to be better than SMT
- Human labor not reusable

Statistical MT

- Requires little human expertise
- Cheap
- Fast
- Good



-- if parallel corpora exist

Electronic Parallel Corpora are on the Rise

Millions of sentence pairs: Arabic-English, French-English, ..

Parliamentary proceedings, UN proceedings, newspapers, ..

Statistical Machine Translation: How

Translation Dictionary

Say we need to translate a French sentence to English:

il croit

Look up the words in a French-English dictionary:

il → he / it

croit → thinks / grows

il — he / it

croit — thinks / grows

il croit

it thinks	bad
he grows	bad
it grows	ok
he thinks	better

Without knowing French, I can say "he thinks"
Is better. Why?

Because "he thinks" occurs more frequently in English text than the other choices!

What is more probable?:

$$
\begin{tikzcd}[row sep=small]
 & \text{Kennedy} \\
\text{John F.} \arrow[ur] \arrow[dr] & \\
 & \text{pencil}
\end{tikzcd}
$$

P(next = Kennedy | John F.) > P(next = pencil | John F.)

0.49 2e-07

Language Model

$$\text{I am} \begin{cases} \xrightarrow{0.02} \text{going} \\ \xrightarrow{7e\text{-}06} \text{went} \end{cases}$$

$$\boxed{P(\text{word} \mid \text{context})}$$

Context is usually the most recent two words.

Use Chain Rule to assign probability to a sentence:

$$\begin{aligned} P(W_1\ W_2\ W_3 \ldots W_n) &= P(W_1)\, P(W_2|W_1)\, P(W_3|W_2\ W_1) \ldots \\ &= \Pi_k\ P(W_k|W_{k-1}\ W_{k-2.} \ldots W_1) \\ &= \Pi_k\ P(W_k|W_{k-1}\ W_{k-2.}) \end{aligned}$$

il croit		LM Probability
	it thinks	3.39e-08
	he grows	7.17e-09
	it grows	3.08e-08
	he thinks	2.33e-07

Our best guess *so far*: il croit = he thinks

Recap: word-for-word translation, using French word order

But, red dress = robe rouge

Word order can be different between source and target!

So let's try again with a different word order:

il croit		LM Probability
	thinks it	5.0e-08
	grows he	1.2e-10
	grows it	4.0e-10
	thinks he	3.9e-08

il croit = he thinks 2.33e-07

So, in this context:

il => he
croit => thinks

Language Model is a powerful tool that does:

- Word sense disambiguation
- Word reordering

Power of Language Model: another example

s nrn stck clmbd nd wll strt ws stll prmtng t ,
a grp f 29 nrn xctvs nd drctrs bgn t sll thr shrs .

stck: stack, stick, stock, stuck
t: to it at out too auto eat tie tea ate toe tee oat iota..

Expected error rate on automatic vowelization in news domain? 5%

Vowelization by LM:

is norian stock climbed and wall street was still promoting it ,
a group of 29 narain executives and directors began to sell their shares .

How to Build Translation Dictionaries?

Parallel Corpus:

That's *my* car
C'est ma voiture

That's *my* brother
C'est mon frere

This is *my* hand
C'est ma main

Co-occurences are the key!

my	C'est	3
my	ma	2
my	voiture	1
my	mon	1
my	frere	1
my	main	1

hand	C'est	1
hand	ma	1
hand	main	1

Co-occurrence => possible translation

Co-occurrence counts => translation probability

P(ma | my) = ?
P(main | my) = ?

If P(y|x) is too small, we say y is not a translation of x.

Result of counting co-occurrences is a statistical dictionary

Statistical Dictionary: known alignments

Suppose a bilingual expert gives us *manual* "alignments":

That's *my* car

C'est ma voiture

This is *my* hand

C'est ma main

That's *my* brother

C'est mon frere

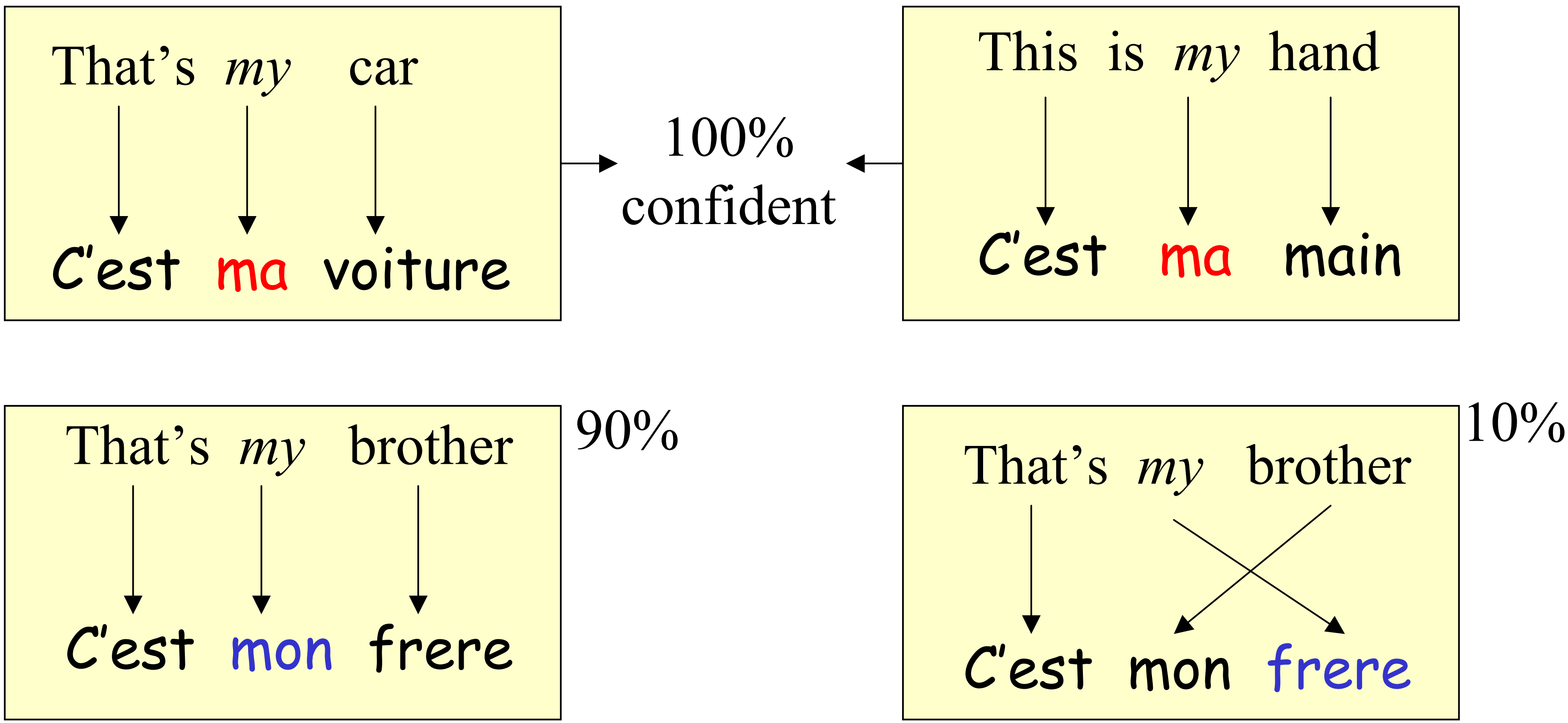
$$P(\text{ma} \mid \text{my}) = \frac{2}{2+1}$$

$$P(\text{mon} \mid \text{my}) = \frac{1}{2+1}$$

Translation probabilities are simply ratios of observed counts!

Statistical Dictionary: known alignments

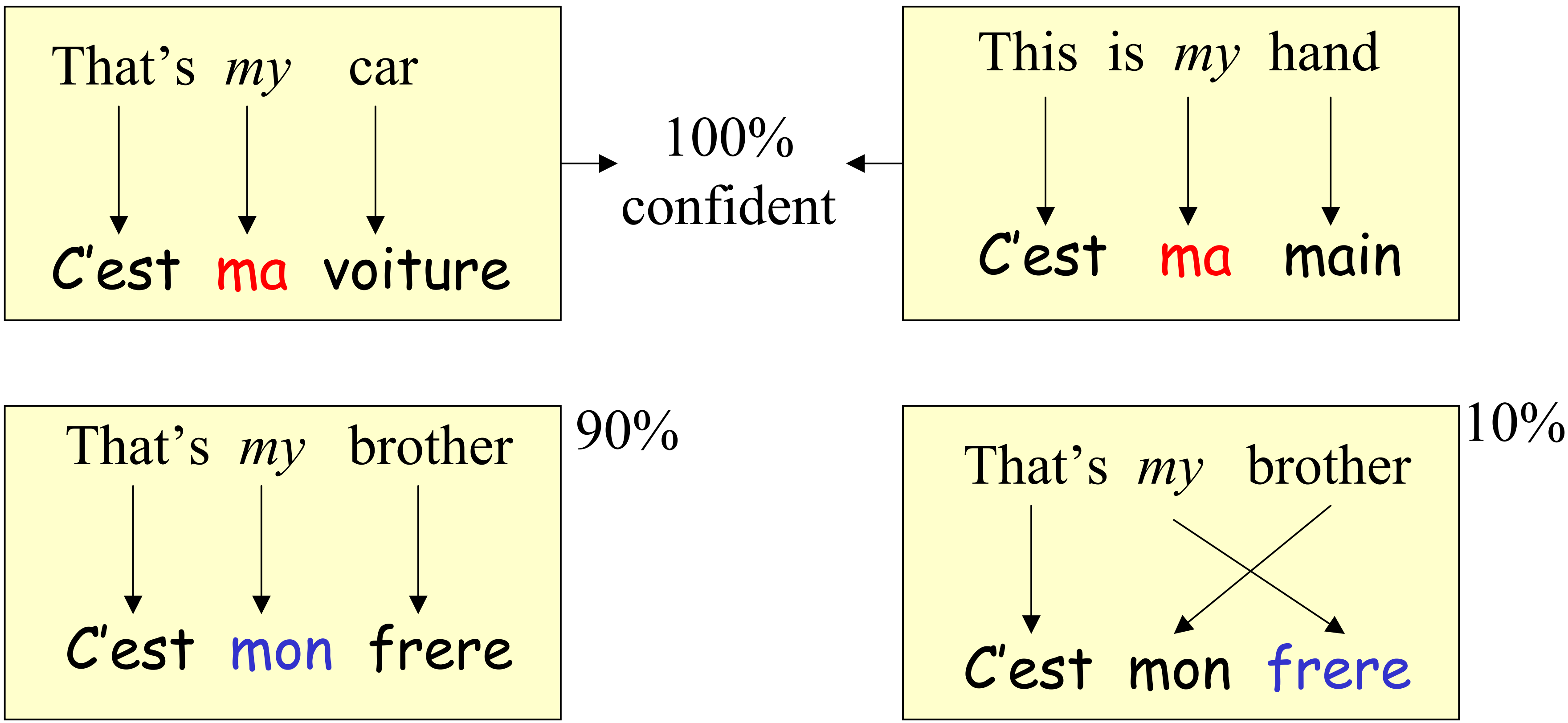
Suppose a bilingual ~~expert~~ provides **uncertain** alignments:



Translation probs are simply ratios of observed **fractional** counts!

my	is connected to **ma**	2	times
my	is connected to **mon**	0.9	times
my	is connected to **frere**	0.1	times

$$P(\text{mon} \mid my) = \frac{0.9}{2+0.9+0.1}$$



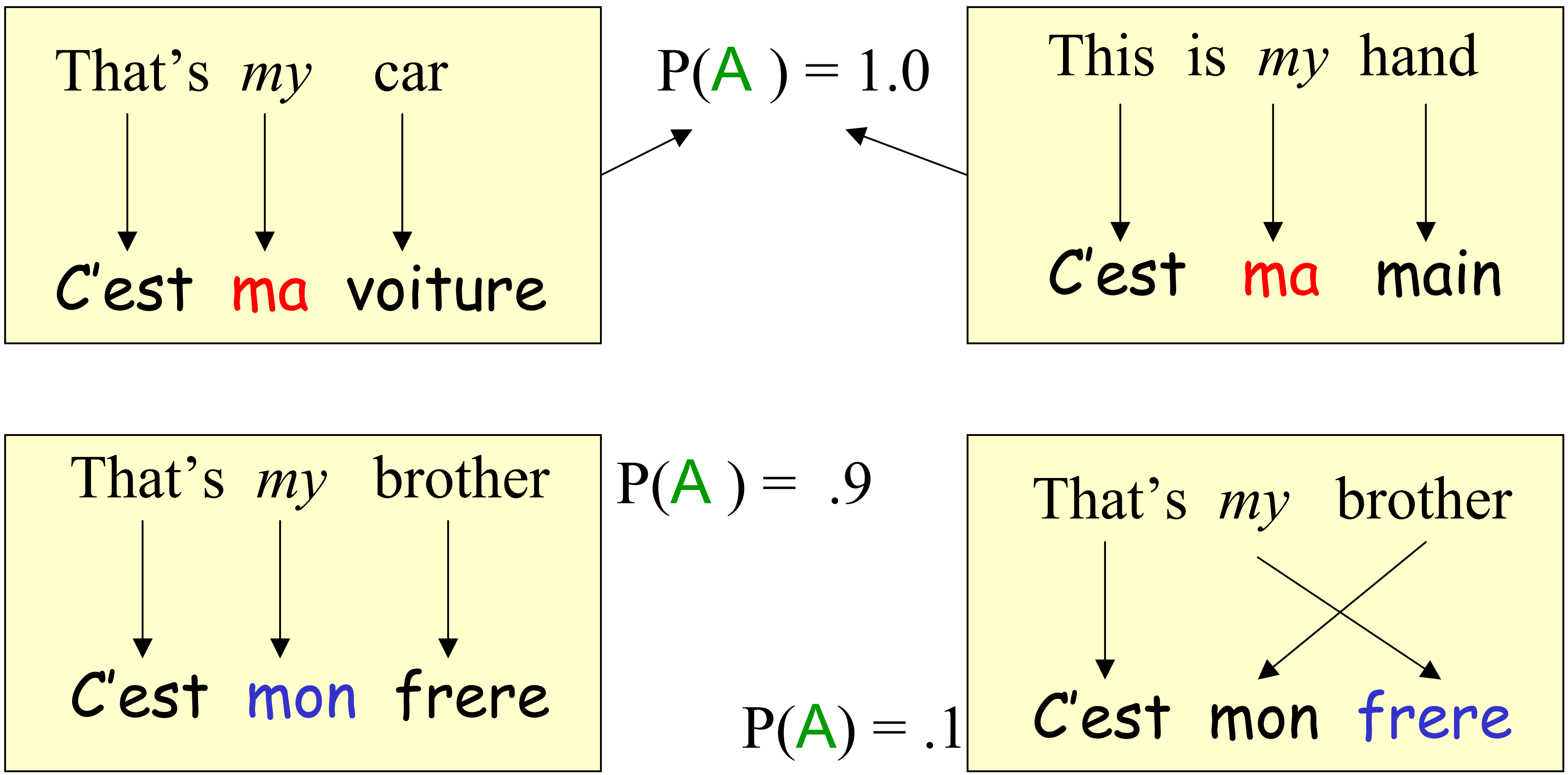
Translation probs are simply ratios of observed **fractional** counts!

$\text{Count}(my, \text{ma}) = 1 \times 1 + 1 \times 1$

$\text{Count}(my, \text{mon}) = 1 \times 0.9$

$\text{Count}(my, \text{frere}) = 1 \times 0.1$

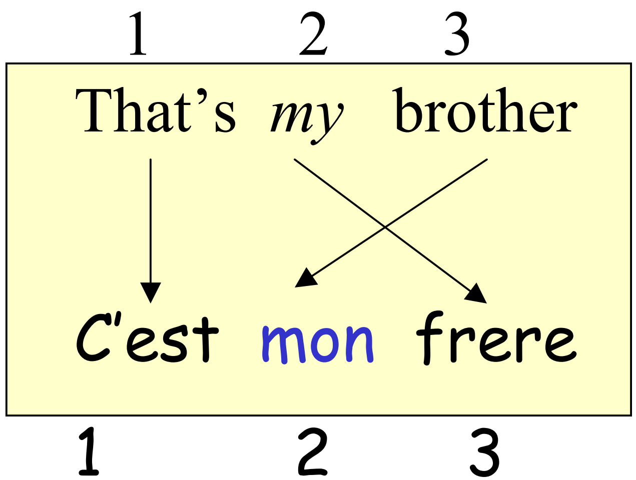
$$P(\text{mon} \mid my) = \frac{0.9}{2+0.9+0.1}$$



P(A) is Alignment Probability.

Alignment Notation

For every French position, remember English position:



F1 => E1
F2 => E3
F3 => E2

More compactly: alignment A $= (a_1, a_2, a_3) = (1, 3, 2)$

Given alignments with their probabilities, we can compute word-to-word translation probabilities!

But it is very expensive to get manual alignments!

We should assume that alignments are **not** given.

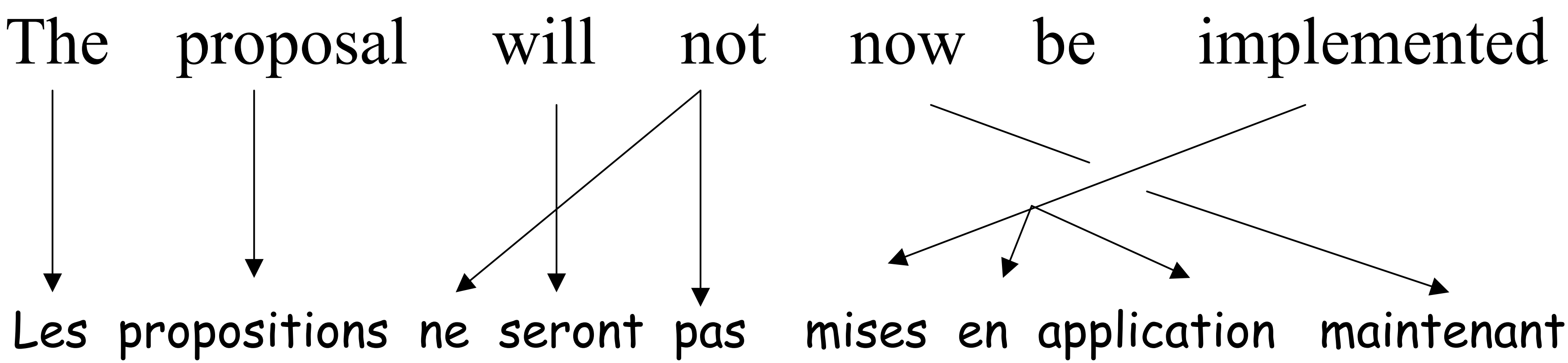
Alignments are **hidden**!

Consider **all** alignments (with some) restrictions

Assign probabilities to alignments

Alignment Restrictions:

1. Word-to-word alignments; not phrase-to-phrase
2. A French word cannot align to multiple E-words



Recap:

Given alignments with their probs, can compute word-to-word translation probs.

We know what the possible alignments are.

Just need to assign probabilities to alignments.

Claim:

Given word-to-word probs, can assign probs to alignments!

Alignment Probability

Given

- French sentence F,
- English sentence E,
- Alignment A
- Word-to-word translation probs $P(f|e)$

how to compute $P(A \mid F, E)$?

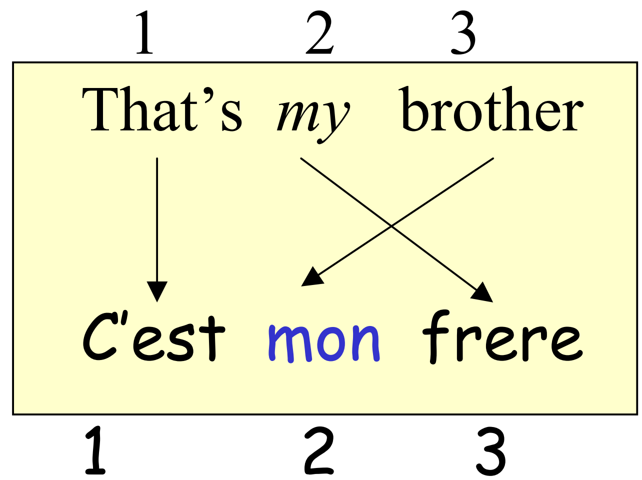
$$P(A, F \mid E) = P(F \mid E)\ P(A \mid F, E)$$

$$P(A \mid F, E) = \frac{P(A, F \mid E)}{P(F \mid E)}$$

But, $P(F \mid E) = \sum_{A} P(A, F \mid E)$

All we need to know is how to compute $P(A, F \mid E)$

$P(A, F \mid E)$



F is a sentence: $f_1 \; f_2 \ldots f_m$

E is a sentence: $e_1 \; e_2 \ldots e_n$

A is alignment: $a_1 \; a_2 \ldots a_m$

$$P(A, F \mid E) = P(a_1 \, a_2 \ldots a_m, f_1 \, f_2 \ldots f_m \mid E)$$

$$= P(a_1) \, P(a_2 | a_1) \ldots P(a_m | a_1^{m-1}) \;\; \times$$

$$P(f_1 \mid A, E) \, P(f_2 \mid f_1, A, E) \ldots P(f_m \mid f_1^{m-1}, A, E)$$

$$P(f_2 \mid f_1, A, E) = P(\text{mon} \mid \text{C'est}, (1,3,2), E) \;\; \approx \;\; P(\text{mon} | \text{brother})$$

$$P(A, F \mid E) = P(a_1\, a_2 \ldots a_m, f_1\, f_2 \ldots f_m \mid E)$$

$$= P(a_1)\, P(a_2|a_1) \ldots P(a_m|a_1^{m-1}) \;\; x$$

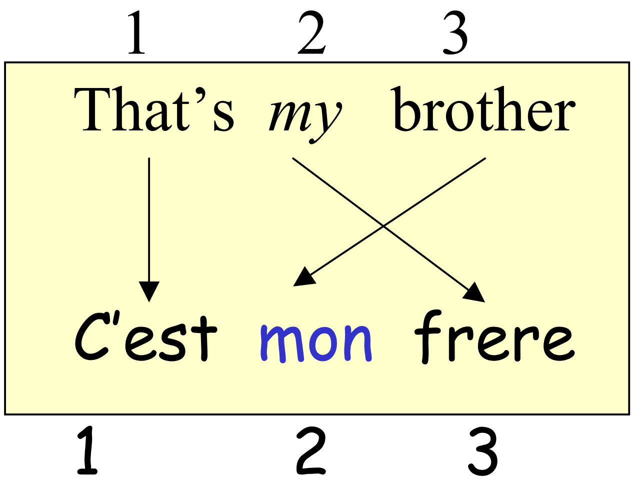
$$P(f_1 \mid A, E)\, P(f_2 \mid f_1, A, E) \ldots P(f_m \mid f_1^{m-1}, A, E)$$

$$\approx \prod_j P(a_j \mid a_1^{j-1}) \;\; x \;\; \prod_j P(f_j \mid e_{a_j})$$

$$\approx n^{-m} \prod P(f_j \mid e_{a_j})$$

with the simplifying assumption: $P(a_j \mid a_1^{j-1}) = 1/n$

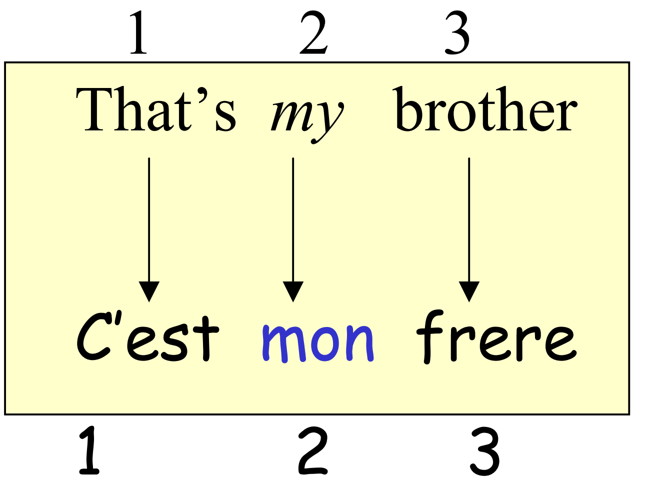
P(A, F | E)



P(A=(1,3,2), C'est mon frere | That's my brother)

= 1/27 x P(C'est | That's) x P(mon | brother) x P(frere | my)

P(A, F | E)



P(A=(1,2,3), C'est mon frere | That's my brother)

= 1/27 x P(C'est | That's) x P(mon | my) x P(frere | brother)

Given w-2-w probs, we now know how to compute P(A,F| E) for any A

$P(A \mid F, E)$

Given w-2-w probs, we know how to compute $P(A,F|\, E)$ for any A

So, we can also compute $P(F \mid E) = \sum_{A} P(A, F \mid E)$

From which we can compute $P(A \mid F, E) = \dfrac{P(A, F \mid E)}{P(F \mid E)}$

If we know w-2-w probs, we can compute alignment probs.

If we know alignments with their probs, we can compute w-2-w probs

Chicken and Egg problem?

1. Start with uniform word-to-word probs.

2. Compute alignment probs using word-to-word probs

3. Compute word-to-word probs using alignment probs

4. Repeat Steps 2-3 until no movement

Can be shown to converge to the optimal solution!

Seed word-to-word probs: example

Parallel Corpus:

That's *my* car
C'est ma voiture

That's *my* brother
C'est mon frere

This is *my* hand
C'est ma main

P(C'est | my) = 1/6
P(ma | my) = 1/6
P(voiture | my) = 1/6
P(mon | my) = 1/6
P(frere | my) = 1/6
P(main | my) = 1/6

P(C'est | hand) = 1/3
P(ma | hand) = 1/3
P(main | hand) = 1/3

Statistical Dictionary: example entry

□ P(□ | English)

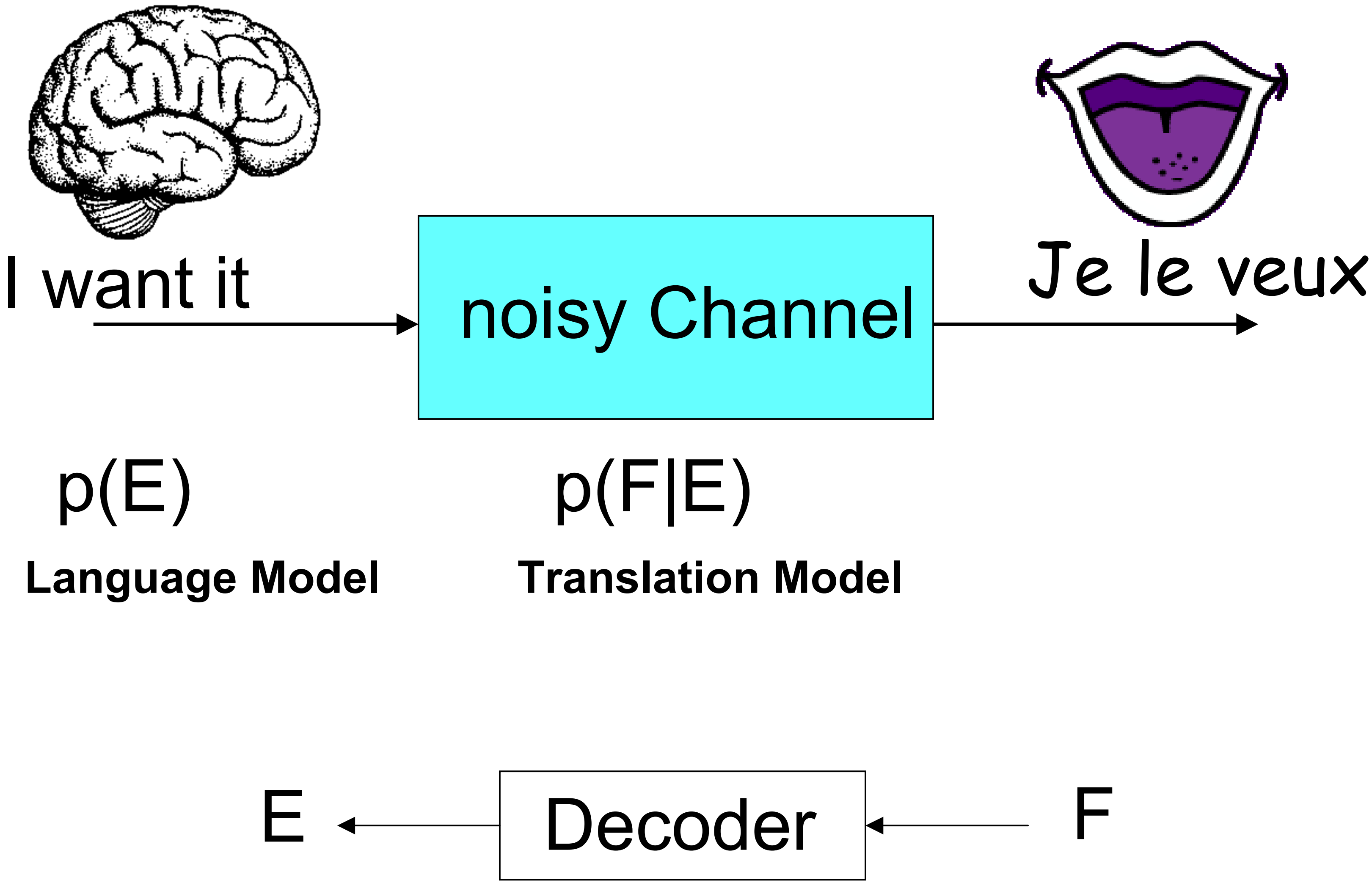
high (0.63), height (0.4), supreme (0.38),
kaohsiung (0.36), tall (0.35), higher (0.34),
antiaircraft (0.33), high-level (0.33), gao (0.31),
highest (0.3), maximum (0.29), hi-tech (0.28),
high-tech (0.28), glad (0.27), high-profile (0.27),
high-speed (0.26), aloft (0.25), raising (0.25),
noble (0.25), raise (0.25), high-performance (0.25),
lofty (0.23), plateau (0.23), senior (0.23),
high-quality (0.22), pleased (0.21), highly (0.21),
elevation (0.21), altitude (0.2), sublime (0.19),
golf (0.18), happy (0.17), expressway (0.15),
new-technology (0.14), upgrade (0.13), elevated (0.12),
hai'nan (0.12), happily (0.12), efficiency (0.1),
enhance (0.1), pleasure (0.1), efficient (0.1), …

Recap:

Learned, by simple counting, how to

- build a statistical dictionary, $P(f|e)$
- write $P(F|E)$ in terms of $P(f|e)$ and alignments
- write $P(E)$ using Language Model

Source-Channel Model



Decoding Foreign

$$\hat{E} \longleftarrow \boxed{\operatorname*{argmax}_{E} \; p(F|E)\, p(E)} \longleftarrow F$$

Bayesian view: $P(E|F) = P(E, F) / P(F)$

$= P(F|E)\, P(E) / P(F)$

Search over all possible English strings?
Efficient decoding is a tough problem.

Translation Model

Sentence-to-sentence probabilities wanted.

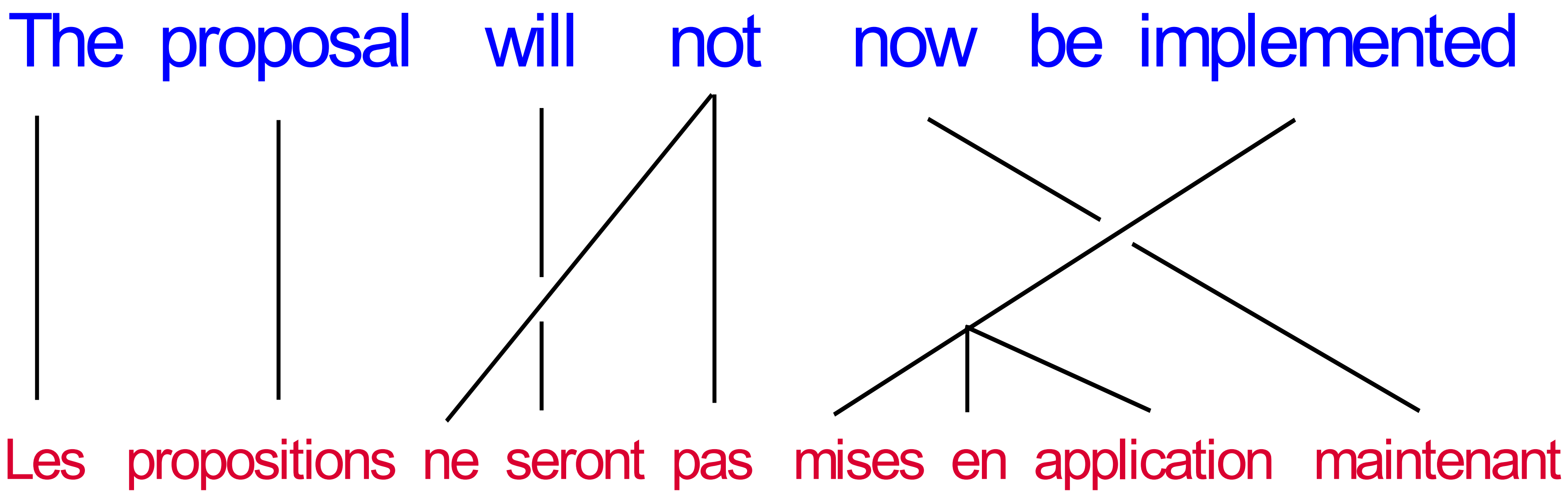
p(Je le veux | I want it) = ?

Decompose into word-to-word probabilities.

But which word goes to which word?

Key Idea: a hidden alignment structure.

Alignments



Given alignment A, can compute p(A, F|E)

But A not given! Sum over all possible A.

Learn p(word|word) and the alignment probabilities from parallel corpora of human translations.

Statistical dictionary + Alignment structure

$\Downarrow$

Channel Model: P(F|E)

Decoding by Dynamic Programming

Je le veux

We do not know in which order these words appear in the translation. (Answer: 1 3 2)

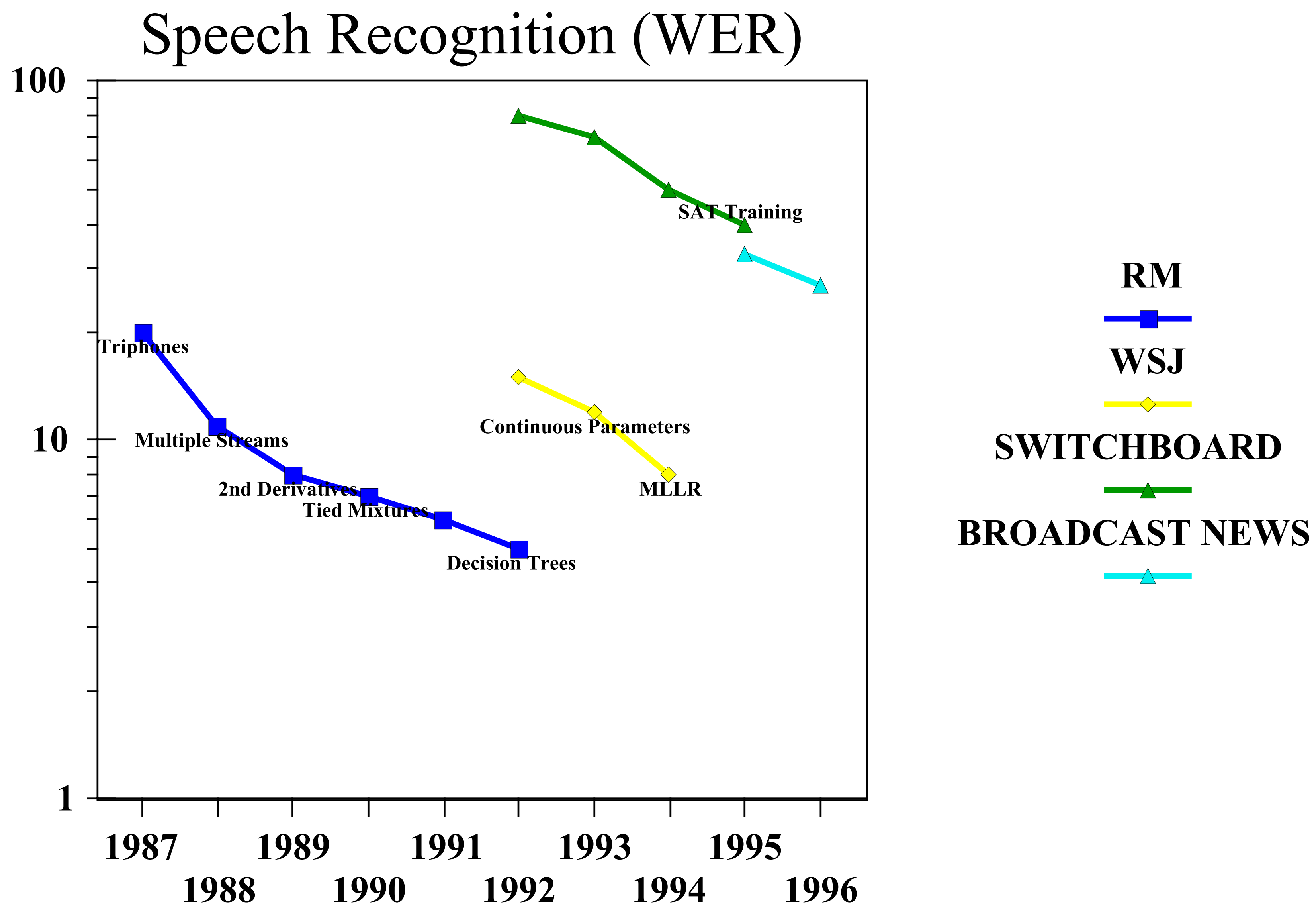
But we should “visit” each word exactly once and translate the word.

Analogous to the Traveling Saleman Problem.
100 words/sec with pruning.

A New Era for Machine Translation

- Large parallel text collections
- Vast computing power
- Reliable automatic metrics

Single number evaluations help drive progress



Human Evaluation: the Ultimate Standard

Expert judges consider many subtle aspects:

- Adequacy
- Fluency
- Grammar
- Idiom
- Style
- …

But human evaluation is expensive, not reusable!

Difficulty of Automatic Evaluation of MT

There is no single ground truth!

There are many correct translations: with genuine word-choice and word-order differences

BLEU (<u>Bi</u><u>L</u>ingual <u>E</u>val <u>U</u>nderstudy) Method

- Goal: automatic metric that approximates human judgment

- Idea:
  - Compare MT to human reference translations
  - Accomodate many gold standards
  - Accomodate word-choice and word-order differences

- Inspiration:
  - Precision & Recall in IR
  - WER in Speech

Many Gold Standards

Ref1: It is a guide to action that ensures that the military will forever heed Party commands .

Ref2: It is the guiding principle which guarantees the military forces always being under the command of the Party .

Ref3: It is the practical guide for the army always to heed the directions of the party .

MT-1: It is a guide to action which ensures that the military always obeys the commands of the party .

(better or worse than?)

MT-2: It is to insure the troops forever hearing the activity guidebook that party direct .

How to judge MT quality?

Words: Count 1-grams in common

Phrases: “look after” => 2-grams
Idioms: “high and dry” => 2,3-grams
Fluency: Count 4-grams in common, etc.

Modified Precision

Reference1: the_1 cat is on the_2 mat

Reference2: there is a cat on the_1 mat

MT: the_1 the_2 the_3 the_4 the_5

Traditional 1-gram Precision = 5/5

Modified 1-gram Precision = 2/5

Similarly for higher-order n-grams

Reference1: the cat is on the mat

Reference2: there is a cat on the mat

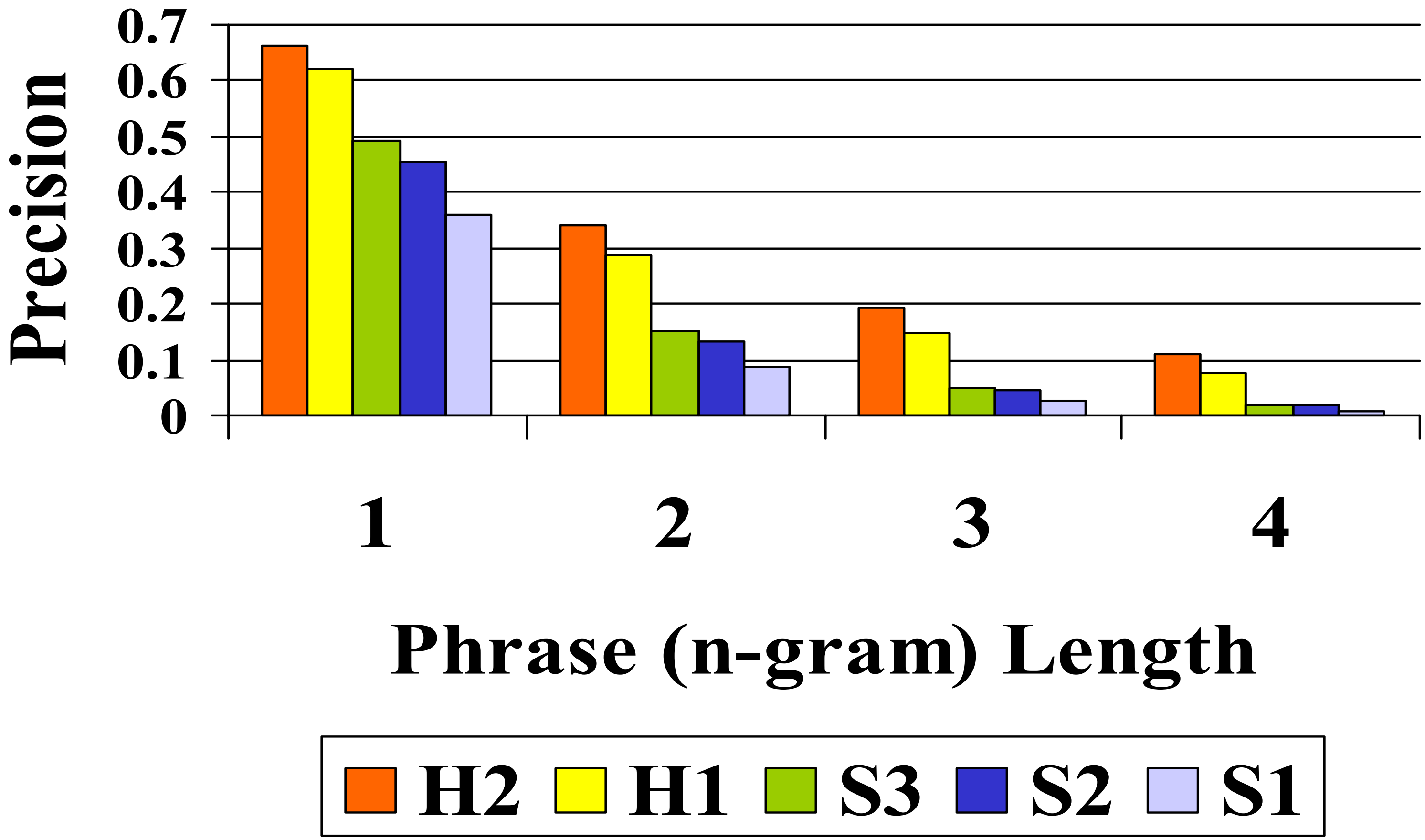
MT -1: **there is a cat on the mat**

MT -2: **that is good**

3g M-Precision, Candidate 1 = 5/5

3g M-Precision, Candidate 2 = 0/1

M-Precision tracks human ranking of translations



Human judgments: H2 > H1 > S3 > S2 > S1

Combining n-gram M-Precisions

Should we combine or just pick one of them?

$$\text{Precision-score} = \exp(W_1 \log P_1 + W_2 \log P_2 + W_3 \log P_3 + W_4 \log P_4)$$

Combining n-gram M-Precisions

Should we combine or just pick one of them?

$$\text{Precision-score} = \exp\left(\tfrac{1}{4} \log P_1 + \tfrac{1}{4} \log P_2 + \tfrac{1}{4} \log P_3 + \tfrac{1}{4} \log P_4 \right)$$

The Flip-side of Precision: Recall

□□□□□□□□□□□□□□□□□□□□□□

The .

Unigram precision = 1.0!

Can get high precision by producing common phrases:

he said ,

Don’t need to know Chinese to see that this is a bad translation!

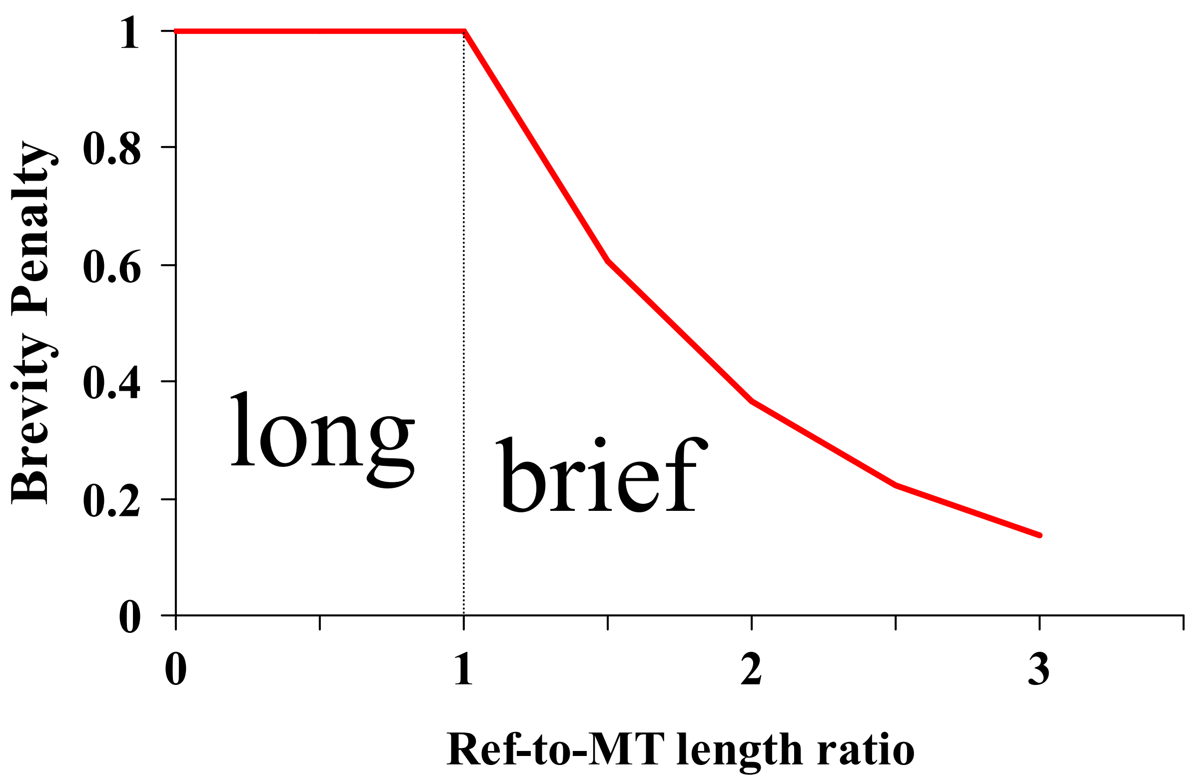
Recall with multiple references

Reference1:	I tossed it
Reference2:	I threw it
`MT-1:`	**`I tossed it`**
`MT-2:`	**`I tossed threw it`**

Brevity Penalty

Too brief ?   Penalize it!

Compare length to the closest of reference lengths



BLEU

BLEU = BP x Precision-score

Normalized to be between 0 and 1

Averaging individual judgment errors

Automatic metrics derive their strength from quantity

Unreliable on just one sentence with just one reference

Quantity leads to quality!

Robustness of automatic metrics

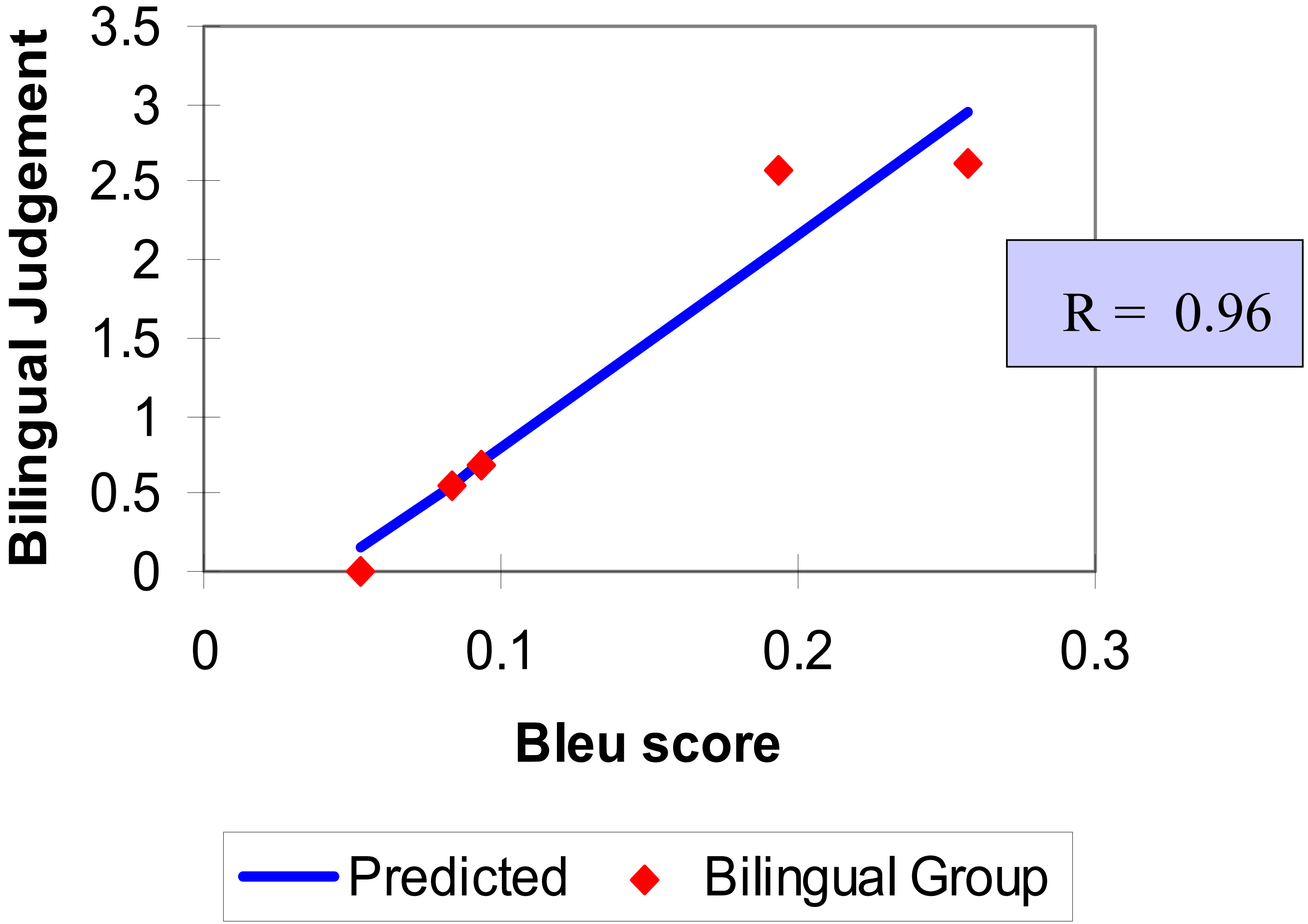
- Across the spectrum of translation quality
- Across language families (HLT'02)

  Arabic → English
  Chinese → English
  French → English
  Spanish → English

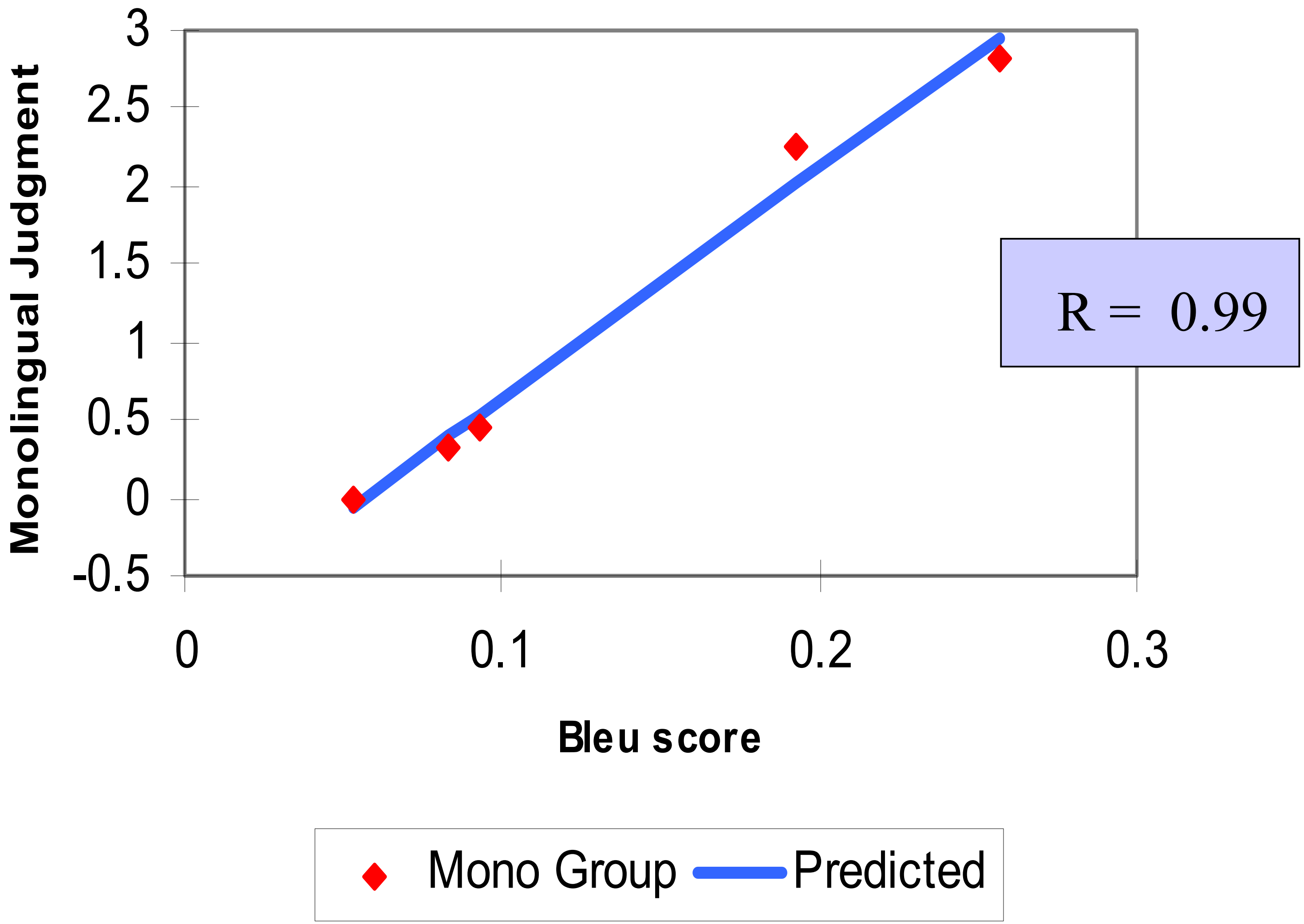
Experimental Set-up: Chinese-English

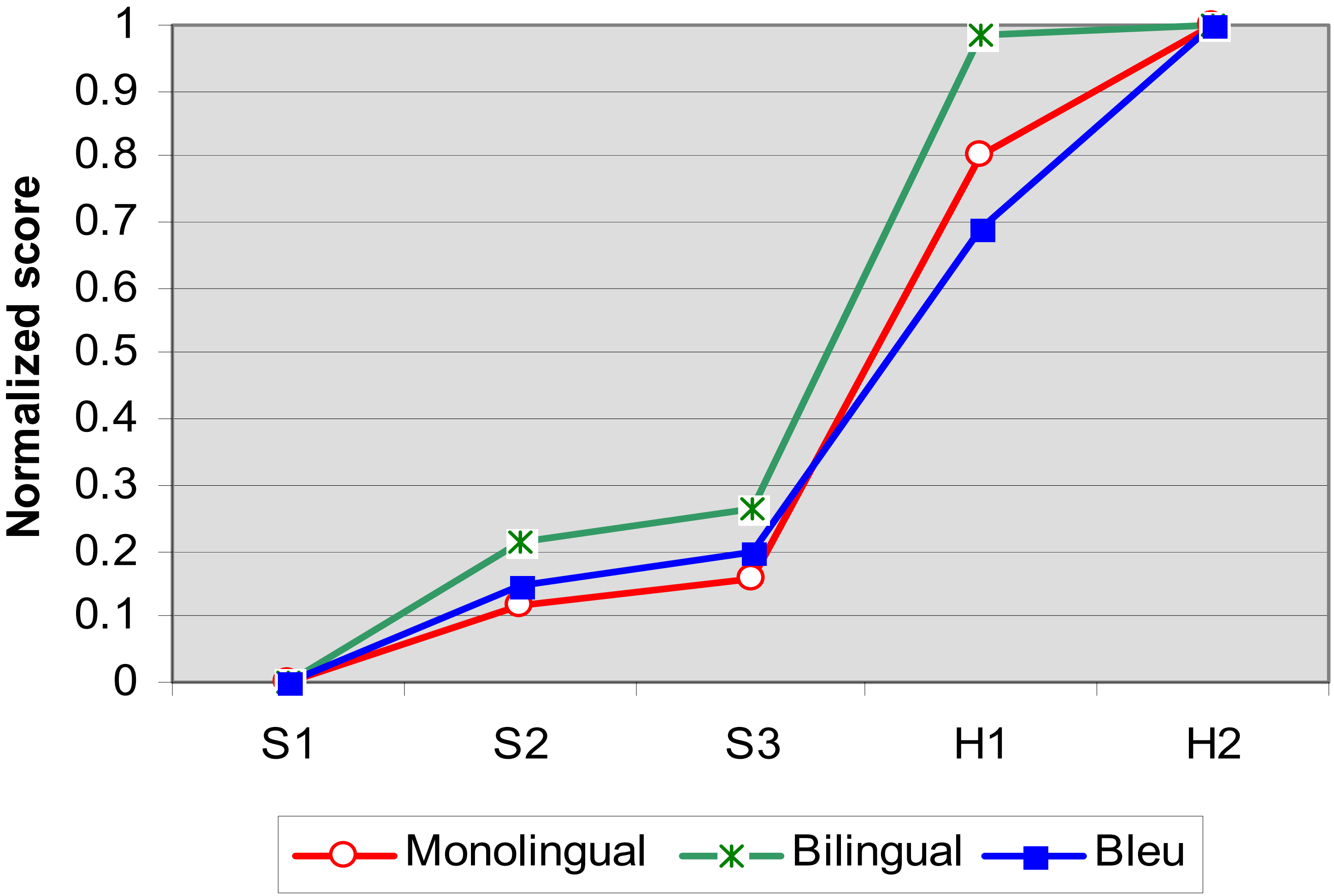
- 40 docs, 2 humans, 3 sys, 2 references
- 15000 words (English)
- Human judges: 10 monolingual, 10 bilingual
- 4500 judgments
- Judge quality from 1 (v. bad) to 5 (v. good)

Pilot Study on Chinese-English Translations



Pilot Study on Chinese-English Translations




Normalized score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
S1
S2
S3
H1
H2
Monolingual
Bilingual
Bleu

Conclusions

Automatic metrics can approximate collective human judgment very well

Vast data, compute power, automatic metrics signal a new era for MT